DOCUMENT RESUME

ED 340 608 SE 052 550

AUTHOR McGinnis, James Randy
TITLE A Comparison of Six Different Diagnostic Procedures Used To Check Raw Quantitative Data for Outliers in a Generic Science Education Study.
PUB DATE 9 Apr 91
NOTE 16p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (Lake Geneva, WI, April 7-10, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computation; *Data Analysis; Educational Research; Elementary Secondary Education; Graphs; Regression (Statistics); Science Education; Statistics
IDENTIFIERS *Outliers

ABSTRACT

This generic science education study demonstrates the contrast of findings obtained through analyzing all the raw data as opposed to removing various combinations of identified potential outliers provided through the application of six diagnostic procedures. Outliers are defined as extreme data points with large residuals. It is argued that an outlier check should be performed on all quantitative data analyzed in science education research to "sound a warning" concerning the accuracy of the collected raw data before standard parametric procedures are applied. Five of the diagnostic procedures are regression diagnostics: the Leverage Point, the Studentized Residual, DFITS, Cooke's Distance, and DFBETA. These values are obtained through the use of Statistical Analysis Software (SAS) by including the influences option in the model step in Proc Reg. A hand calculation of critical values is then done to obtain cut-off values to identify potential outliers. The other diagnostic procedure demonstrated is a visual inspection of the scatter plot of the raw data to identify suspicious data points. Cooke's Distance, being the most conservative procedure, is recommended as the preferred diagnostic procedure to highlight suspicious points in science education research. The outlier diagnostic procedure, the SAS procedure to obtain calculated diagnostic values, and a calculation of diagnostic cut-off values are appended. (KR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED340608

A Comparison Of Six Different Diagnostic Procedures Used to Check Raw Quantitative Data For Outliers In A Generic Science Education Study

James Randy McGinnis
Department of Science Education
The University of Georgia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James Randy McGinnis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual meeting of the
National Association for the Research in Science Teaching
Lake Geneva, Wisconsin, April 9, 1991

BEST COPY AVAILABLE

SE 052 532

Abstract

This methodological study compares six different diagnostic procedures used to check raw quantitative data sets for outliers in a generic science education study. Outliers are defined as extreme data points with large residuals. It is argued that an outlier check should be performed on all quantitative data analyzed in science education research to 'sound a warning' concerning the accuracy of the collected raw data before standard parametric procedures are applied.

Five of the diagnostic procedures demonstrated are regression diagnostics: the Leverage Point, the Studentized Residual, DFITS, Cooke's Distance, and DFBETA. These values are obtained through the use of Statistical Analysis Software (SAS) by including the influences option in the model step in Proc Reg. Hand calculation of critical values then is done to obtain cut-off values to identify potential outliers. The other diagnostic procedure demonstrated is a visual inspection of the scatter plot of the raw data to identify suspicious data points.

This generic science education study demonstrates the contrast of findings obtained through analyzing all the raw data as opposed to removing various combinations of identified potential outliers provided through the application of all six diagnostic procedures. Cooks's Distance, being the most conservative procedure, is recommended as the preferred diagnostic procedure to highlight suspicious points in science education research.

Purpose Of The Study

Science education researchers frequently find it necessary to statistically analyze quantitative data collected in their own studies or brought to them by classroom science teachers who have conducted instructional studies with their students as subjects. In the case of quasi-experimental designs or the rare true experimental designs with more than two groups, it is hoped through the use of inferential statistics (usually the Analysis of Variance, ANOVA, or the General Linear Model, GLM) to show that the treatment introduced made a statistical difference. What these studies have in common is the reliance on a collected raw data set to be mathematically analyzed. The accuracy of the data set collected determines to the greatest extent the worth of the computed result of the statistical procedures selected. Even one very inaccurate data point that has been added to the data set can distort the result of a statistical analysis to cause a study to not show statistical significance when it really was statistically significant. It is argued, therefore, by many statisticians that the data set of every study be the object of a diagnostic check for outliers, influential, inaccurate data points with large residuals, before the data is used for hypothesis testing of a research question (Chatterjee & Price, 1977 ; Hawkins, 1980; Rousseeuw & Leroy, 1987). The goal of the diagnostic check of the data is to identify the outliers and examine the consequences of their presence in and out of the data set.

The purpose of this paper is to compare six different diagnostic procedures used to check raw quantitative data for outliers in a generic science education study. Five procedures are regression diagnostic; the other is the scatter plot. Refer to Appendix A for definitions of each procedure.

Significance

Currently in science education, it not common practice for researchers to statistically analyze raw data for outliers before using inferential statistics to hypothesis test. It is argued that as a result, many results obtained are suspect of undue influence of inaccurate data points. In this generic science education study of an outlier check of data, I illustrate the use of six different diagnostic procedures available on Statistical Analysis Software (SAS) to check data for outliers. Through this comparison of six of the more commonly used outlier diagnostic procedures, it is shown how they statistically differ and how the removal of even one outlier can often drastically change statistical results.

Design And Procedures

Example data statistically analyzed consisted of the records of one-hundred and eleven randomly selected high school students. These students consisted of fifty-five females and forty-six males. All students participated in the instructional science education study and had the same treatment. Eight different

measurements were taken to document pre- and post- scores. Of interest in this outlier study were the data for only two of the measurements (the rationale being that a diagnostic check of one of the variables can serve as an example of the procedure for the remaining ones).

Data were arranged and inputted into a data file in a word processing program. It was decided initially to remove students from the data file who had not provided their gender (one) and who had not recorded both the pre-measure and the post-measurement for the variable selected for study. This reduced the sample to ninety-two students, forty-nine females and forty-three males.

The statistical procedure selected in SAS to analyze the study was the General Linear Model, GLM. The dependent variable was the post-measure. The interaction model of $y = b_0 + b_1$ Pre-measure + b_3 (interaction of pre-measure and gender) was tested with all of the raw data.

Five regression diagnostic procedures to check for outliers in the raw data set were then performed on SAS by using Proc Reg and including the influence option in the model step. Appendix B contains the SAS program. Obtained were the leverage point, the Studentized Residual, DFITS, Cooke's Distance, and DFBETA. Proc Plot was also used to produce a scatter plot of the pre- and post-measures, a sixth visual diagnostic procedure for outliers. Potential outliers were identified by calculating critical values for each regression diagnostic procedure and by identifying

points in the scatter plot away from the other points. Table 1 illustrates these calculations. These potential outliers were then put in table form under the headings of the six diagnostic procedures. These points were also marked on the scatter plot.

The GLM procedure was then repeated first with the removal of the least number of potential outliers identified in a diagnostic procedure (Cooke's Distance, one data point) and then with all possible different outliers removed indicated by all six diagnostic procedures (thirteen data points).

Results

The GLM procedure using all the data indicated an interaction between the pre-measure and gender. The GLM procedure of the interaction model, $y = b_0 + b_1$ Pre-measure $+ b_3$ (interaction of pre-measure and gender), had $F_{3, 92} = 23.45$, $p < .0001$, indicating that the interaction model was statistically significant at $\alpha = .05$. The F-value of the interaction term, Pre-measure*gender, $F_{1,92}$, had a F- value of 4.40, with $p < .0387$. The R-squared value for the model was .4424.

A diagnostic check of the data indicated different potential outliers to examine. The scatter plot indicated 5 possible outliers; the Leverage Point indicated 7 possible outliers (with values greater than the calculated critical value of .0869, $n = 95$ with 3 missing values; the Studentized Residual indicated 9 possible outliers (with values greater than the calculated

critical value of 1.664, n = 95 with 3 missing values; the DFITS indicated 4 possible outliers (with values greater than the calculated critical value of .4170, n = 95 with 2 missing values); Cooke's Distance indicated 1 possible outlier (with a value greater than the calculated critical value of .7000); and the DFBETA indicated 3 possible outliers (with values greater than .2085). No one point was identified by all tests as an outlier. One point was identified as an outlier in five of the six tests; two points were identified as an outlier in four of the tests; one point was identified as an outlier in three of the tests; three points were identified as an outlier in two of the tests; and six points were identified as an outlier in only one of the tests.

The GLM procedure was performed on the data without the one point indicated by Cooke's Distance. This time, the correlation between the pre-measure and gender was .23616, $p < .024$. In this analysis, n = 91, the interaction was not significant. The interaction term, pre-measure*gender, had $F_{1, 91} = .60$, with $p < .4423$. That was not significant at $\alpha = .05$. The better model was the main effects model with $F_{2, 91} = 21.99$, $p < .0001$. Both pre-measure and gender significant at $\alpha = .05$. R-squared was equal to .3323.

The GLM procedure was then performed on the data without the thirteen points indicated by any of the diagnostics procedures. The correlation coefficient between the pre-measure and Gender increased to .35086, $p < .0015$. In this analysis with n = 79,

the interaction term was not significant, $F_{1, 79} = .20$, $p < .6541$. The main effects model was $F_{2,79} = 34.33$, $p < .0001$, but gender was not significant at $F_{1, 79} = 2.74$, $p < .1020$. Therefore, the main effects model with both pre-measure and gender was not appropriate.

Interpretations And Conclusions

The GLM procedure indicated that there was an interaction between the pre-measure and gender when all the raw data set was used. This was seemingly a finding of some note in this generic science education study. For not only was the treatment significant, but it varied in its effect on males and females depending on their pre-measure, a highly interesting occurrence for science educators. Diagnostic procedures, however, identified potential outliers in the data set which had contaminated the result. When the GLM was run again with only one outlier removed (the one obtained through Cooke's Distance), the interaction term was not significant, although the Main Effects model with the pre-measure and gender was. This indicated the treatment still had a significant effect, but not differently on males and females as determined by their pre-measures. However, when the GLM procedure was run with all thirteen possible outliers removed, besides for there being no significant interaction term, even the Main Effects model was not appropriate. This indicated that the model should not have both the pre-measure and gender.

What is shown through this comparison of six diagnostic procedures to identify outliers is that the consequences of including inaccurate collected data points with large residuals can have a significant effect on the outcome of the statistical analysis of the data set in science education. And while it is also shown that the various diagnostic procedures do not with absolute confidence identify the same suspect points, the procedures do serve to highlight measurements which need to be checked for accuracy. Points identified through Cooke's Distance, being the most conservative diagnostic procedure, are particularly suspect. Removing those points from the raw data is recommended unless their accuracy can be guaranteed.

References

Chatterjee, S.& Price, B. (1977). Regression analysis by example. New York: John Wiley & Sons.

Hawkins, D.M. (1980). Identification of outliers. New York: Chapman and Hall.

Rousseeuw, P.J.& Leroy, A.M. (1987). Robust regression and outlier detection. New York: John Wiley & Sons.

Appendix A

Outlier Diagnostic Procedures

1. Scatter plot of all data.

   Evaluation of graph:

   Any point (or points) appreciably away from the rest of the data is a potential outlier.

2. Leverage Point.

   This shows how far the point is from the mean.

   Equations:

$$\hat{h} = \frac{1}{N} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}$$

   Evaluation of values (determination of cut-off value):

   a. For Large N (greater than 50)

   If $\hat{h} > \frac{2(k+1)}{N}$, then it is a potential outlier.

   b. For Small N (less than 50)

   If $\hat{h} > \frac{3(k+1)}{N}$, then it is a potential outlier.

3. Studentized Residual

   Equation:

$$\varepsilon^* = \frac{\varepsilon}{\sqrt{s^2_{(i)} \cdot (1-h)}}$$

   (i)* without that point

   Evaluation of values:

   It is a potential outlier if $\varepsilon^* > t_{n-k-2}$

4. DFITS

Equation:

$DFITS = \sqrt{h_i / 1 - h_i} \; (\varepsilon / \sqrt{s^2}_i)$

Evaluation of values:

It is a potential outlier if $DFITS > 2 \sqrt{k + 1 / N}$

5. Cooke's Distance

Equation:

$D = \varepsilon^2 / (k + 1) \; MS_{reg}$

Evaluation of values:

It is a potential outlier if $D > 1$ or $F_{k+1,\, N-(k+1)}$

6. DFBETA

Equation:

$DFBETA = b_i - b_{(i)^*} / SE_{bi}$

$(i)^*$ without that point

Evaluation of values:

It is a potential outlier if $DFBETA > 2/\sqrt{N}$ or 1

Appendix B

SAS Procedure to Obtain Calculated Diagnostic Values

1. Scatter plot

```
Proc Plot;
     plot y*x;
run;
```

2. Other 5 diagnostic tests: the Leverage Point, the studentized residual, DFITS, Cooke's Distance, and DFBETA

```
Proc Reg simple;
     model y = x/r influence;
     output=new predicted yhat residual=ehat;
run;
```

Author Notes

I gratefully acknowledge Dr. Stephen F. Olejnik in the Department of Educational Psychology at The University of Georgia for introducing me to the concept of outlier diagnostics in EPY 816, Correlation and Multiple Regress, during the summer quarter 1990.

Table 1

Calculation of Diagnostic Cut-Off Values

Example: N = 92, k = 3 (from interaction model)

1. Scatter plot

   Evaluation:

   Visual inspection to identify those points suspiciously away from others.

2. Leverage Pt.

   Evaluation:

   Large $N > 2(3 + 1)/92 = 0.086956521$

3. Studentized Residual

   Evaluation:

   $*t_{n-k-2,\ \alpha}$

   $t_{92-3-2,\ .05} = 1.66666$

4. DFITS

   Evaluation:

   $2\sqrt{3 + 1 / 92} = 0.4170$

5. Cooke's Distance

   Evaluation:

   $D > 1$ or $F_{3+1,\ 92-(3+1)} = 0.738$

6. DFBETA

   Evaluation:

   DFBETA $> 2/\sqrt{92} = .2085$ (more conservative) or 1 (less conservative)